

Les unités de signification spécialisées: élargissant l'objet du travail en terminologie¹

Rosa Estopà Bagot

Institut Universitari de Lingüística Aplicada (Universitat Pompeu Fabra)

Rambla Santa Mònica, 30

08002 Barcelona

rosa.estopa@trad.upf.es

Abstract

In this job we are outlining what must be the basic object of the work in terminology which responds to the needs of the users. Firstly we propose that this object cannot limit itself to noun referential units, that is to say, to the terminological units, it has to include any unit of specialised meaning to the specialised text. Secondly, we defend the right to distinguish between relevant and non-relevant units of specialised meaning for a specific professional purpose. A difference which allows us to formulate user profiles of concrete terminological applications. To argue these statements we are starting from the theoretical foundations of The Communicative Theory of Terminology of M. Teresa Cabré, and we are basing it on two experimental tests which were carried out in the field of term detection.

Key-words: Terminology, Terminological Object, Terminological Applications

1. L'objet du travail en terminologie

La reconnaissance des unités d'un texte avec un signifié spécialisé connu comme dépouillement terminologique est une des étapes fondamentales de tout travail en terminologie (élaboration de vocabulaires, glossaires, bases de données, bases de connaissance, thesaurus, préparation de traductions, indexation de textes, construction de vérificateurs orthographiques, etc.). Cette étape, qui représente la

¹ Je voudrais apporter mes remerciements à Virginie Inda et à Marie-Claude L'Homme pour la supervision de la traduction de ce document. Ce travail provient de la thèse doctorale *Extracció de terminologia: elements per a la construcció d'un SEACUSE* (1999) dirigée par M. Teresa Cabré.

tâche centrale de tout travail terminologique, est loin d'être simple. En effet, elle requiert beaucoup de temps et de systématisation au moment d'appliquer les critères de sélection de termes. C'est une des étapes les plus contraignantes et longues surtout lorsque l'on manipule des volumes d'informations importantes. Le principal risque encouru est d'aboutir à un travail peu systématique et, par conséquent, peu efficace. Depuis quelque temps, les terminologues peuvent avoir recours, pour cette étape de leur travail, à des extracteurs automatiques de termes qui leur proposent des listes de candidats à termes qu'ils doivent épurer.

Dans le cadre de la Théorie Communicative de la Terminologie (Cabré, 1999, 2000), le dépouillement terminologique est conditionné par le principe d'adéquation qui permet d'adapter les produits terminologiques aux destinataires et à leurs besoins. Ainsi, tout dépouillement terminologique est guidé par des questions comme : que doit-on sélectionner? d'où ? pour qui ? dans quel but ?

Mais, si nous faisons une révision des unités qu'on inclue dans la plupart des travaux terminologiques (vocabulaires, dictionnaires, bases de données, thesaurus, etc.), nous réalisons que dans la majorité des cas, leurs entrées sont exclusivement des noms, c'est à dire des unités terminologiques. Seules très peu de travaux incluent aussi quelques verbes et quelques adjectifs, et plus rarement quelques unités phraséologiques. Pourquoi ? Quel doit être l'objet de travail des applications terminologiques ?

Dans ce article, en premier lieu, nous nous centrerons sur le thème que représente l'objet de sélection , c'est à dire, un des conditionnements qui favorise l'application terminologique; et, par la suite, nous nous attacherons aux besoins que différents usagers ont dans leurs activités professionnelles.

Nous montrerons, d'une part, que les unités à considérer ne devraient pas se limiter aux seules unités nominales et, d'autre part, que le nombre ainsi que le type d'unités à extraire dépend du type d'application pour laquelle elles sont relevées (transmission de la connaissance, documentation, traduction ou terminographie). Nos observations s'appuient sur deux expérimentations que nous avons menées et qui sont décrites plus loin. Enfin, nous présenterons comment ces considérations peuvent être prises en compte par les extracteurs automatiques de terminologie.

2. L'objet d'extraction: les USS dans les sciences de la santé

L'unité terminologique (UT), comprise dans le sens d'unité lexicale nominale du langage naturel a été considérée comme l'unité de base de la terminologie. Mais, à la fin du siècle dernier, plusieurs auteurs ont repensé l'objet de la terminologie et ou ils ont élargi la notion d'unité terminologique ou encore ils ont proposé une autre unité plus large qui inclurait d'autres unités que les unités lexicales de caractère nominal².

Dans une perspective plus pratique de la terminologie, il est clair que la plupart des applications incluent habituellement des unités terminologiques, de catégorie grammaticale nominale; certaines applications incluent intègrent aussi d'autres unités et d'autres catégories grammaticales. On constate, par exemple, que les grandes banques de données faites pour des traducteurs ont eu besoin d'incorporer, aux côtés des UT, des unités phaséologiques. Or les vocabulaires élaborés dans le cadre d'une politique linguistique (au Québec, en Catalogne, au Pays Basque) incluent quelques unités lexicales verbales ou adjectivales, même si ce n'est pas systématiquement.

Depuis ces vingt dernières années, les applications terminologiques, dans le cadre du génie linguistique ou des industries de la langue, se sont multipliées et diversifiées. Dans ce champ, un des thèmes des les plus travaillés est l'extraction automatique de terminologie. Comme nous l'avons expliqué lors du dernier point, au cours du processus de tout travail terminographique, la tâche considérée la plus lourde et celle qui demande un plus vaste investissement de ressources et de temps est le repérage de termes à partir de textes. Jusqu'à présent, cette opération s'est faite à la main, c'est à dire grâce à la reconnaissance, une par une, de toutes les unités terminologiquement pertinentes dans un texte.

Avec la fonction d'automatiser l'étape de dépouillement du travail terminologique pour gagner en rapidité et en systématisation, la fin des années quatre-vingt voit apparaître, au Canada, le premier extracteur automatique de terminologie [TERMINO, 1988, (Perron, 1989 ; David et al., 1991)]. Cependant aujourd'hui la recherche sur les extracteurs automatiques de terminologie est à la pointe de l'actualité et il existe déjà plusieurs extracteurs sur le marché, comme [Acabit (Daille, 1994, 1996)], [LEXTER (Bourigault, 1993, 1994)], [FASTR (Jacquemin, 1996)], [Neural (Frantzi et al., 1995)], (Heid et al., 1996) etc. Ils ont été surtout conçus pour reconnaître des termes à partir des unités terminologiques

² Bien qu'ils ne nient pas quelques d'autres unités spécialisées, certains auteurs comme Wüster (1985 (1998)), Rey (1979), Dubuc (1985), Sager (1990) ou Lérat (1995) réservent le mot UT à une dénomination ou à unité lexicale nominale; alors que d'autres comme Cabré (1992), Auger et Rousseau (1978), on considéré que la meme UT inclue également d'autres categories gramaticales.

polylexicales, car ces unités sont les plus fréquentes du discours spécialisé, principalement dans les domaines techniques, et elles montrent de nombreux traits formels de reconnaissance.

Mais depuis le début, plusieurs études ont démontré que une des limites récurrentes des extracteurs est qu'ils sont très restrictifs à ce qu'ils retiennent, de manière qu'ils se concentrent sur la détection des UT polylexicales (UTP), de catégorie grammaticale nominale. Quelques travaux qui ont évalué des logiciels d'extraction corroborent cette constatation (Kagueura et Umino, 1996), (Drouin, 1997), (Cabré, Estopà, Vivaldi, 2000) (Amar et David, 2001).

De toute évidence, ces dernières sont les unités les plus prototypiques et les plus fréquentes des textes spécialisés, elles sont aussi celles qui présentent les caractéristiques morphosyntaxiques les plus explicites, ce qui facilite leur extraction. Cette restriction a impliqué une réduction de la reconnaissance automatique à un type de mots: les UT et, dans de cette catégorie, à un seul type de structure: les unités syntagmatiques (Voutilanen, 1993), (Lauriston, 1994), (David, 1995), (Naulleau, 1998). Cette restriction des logiciels d'extraction basés sur la connaissance linguistique se justifie avec des arguments comme les suivants:

- La plupart des UT d'un domaine de spécialisation sont des unités syntagmatiques et, pour cette raison, il est inutile de compliquer le système en lui faisant reconnaître les termes simples.
- Formellement, les unités monolexicales spécialisées et les générales ne se distinguent pas. Pour cette raison, il est impossible de les différencier.
- Les unités monolexicales sont beaucoup plus polysémiques que les polylexicales et, pour cela, travailler avec les heuristiques sémantiques rapportées aux termes monolexicaux est beaucoup plus compliqué que ne pas introduire de connaissance sémantique dans les UTP.

Enfin, il est vrai que d'un certain point de vue morphosyntaxique, les UTP sont plus faciles à détecter que les unités monolexicales, étant donné qu'elles présentent une structure morphosyntaxique explicite contrôlable. Mais malgré cela, les trois arguments doivent être nuancés.

En relation avec le premier argument, on considère que près de 80% des terminologies sont formées par des UTP. Il convient toutefois de préciser que ces chiffres sont basés sur les unités codifiées en répertoires lexicaux et terminologiques, mais s'ils partent de l'usage réel des termes dans les textes, de une approximation textuel à la terminologie (Slodzian, 1995) (Cabré, 1999), ces affirmations ne seraient pas tout à fait valides. Des tests réalisés dans le domaine de la biomédecine signalent que l'ensemble des

unités monolexicales spécialisées d'un texte thématiquement spécialisé ne peut être ignoré car cela correspond au 35% et 45% des unités avec un signifié spécialisé.

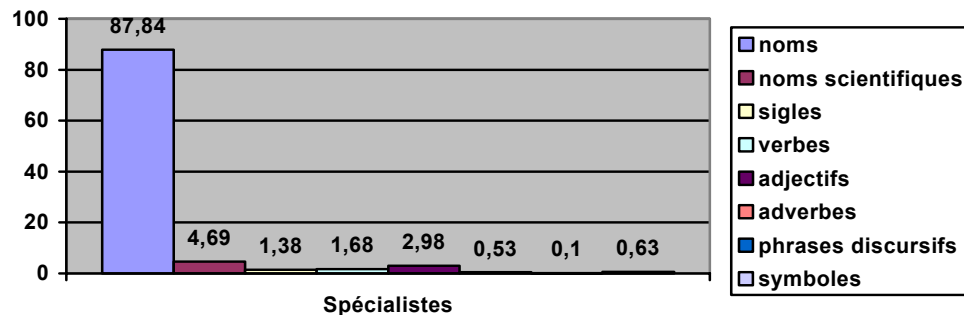
Les trois arguments sont aussi discutables en raison de l'importance accordée à la catégorie nominale. Ainsi, il est vrai que les UT sont toujours nominales et qu'elles sont les USS les plus fréquentes des textes spécialisés; mais il est également certain que les verbes, les adjectifs et les adverbes avec un usage thématiquement spécialisé ont aussi un rôle dans ce type de textes.

Pour ces raisons, nous pensons que ces affirmations pourraient être reconsidérées parce que, bien qu'il soit certain que les unités monolexicales simples sont assez idiosyncratiques et souvent polysémiques (et, par conséquent, il est difficile de discriminer le sens spécialisé ou général d'une unité simple sur la base de critères linguistiques), il y a d'autres unités monolexicales construites —dérivés, composés, abrégés— qui présentent des particularités formelles. Les extracteurs pourraient ainsi se baser sur ces dernières pour les détecter.

Il est donc nécessaire de repenser l'objet de base du travail en terminologie, et plus concrètement nous nous sommes interrogés quant à l'objet des extracteurs. Nous nous sommes basés sur le principe d'adéquation postulé pour la Théorie Communicative de la Terminologie (Cabré, 1999). Selon ce principe chaque application doit s'adapter aux caractéristiques de son usage (à leurs objectifs, leurs contextes, leurs destinataires, leurs finalités professionnelles, etc.). Premièrement, pour avaliser l'idée que l'objet d'une application terminologique doit aller au-delà de l'UT, nous avons réalisé un test expérimental. Ce test a eu comme but l'analyse des dépouillements manuels faits par trois spécialistes. On a choisi des spécialistes parce que, conventionnellement, on les considère comme les professionnels idéaux pour reconnaître les UT d'un texte car ils possèdent la compétence cognitive et pragmatique la plus haute dans leur spécialité.

Dans notre cas, nous avons travaillé avec un corpus textuel du domaine de la biomédecine, où nous avons analysé les dépouillements manuels qu'ont réalisés des docteurs spécialistes en médecine interne. Plus précisément, ces professionnels ont dépouillé le texte "Enfermedades infecciosas por Rickettsia" (qui fait parti du chapitre sur les Enfermedades infecciosas du livre Medicina interna (Farreras et Rozman, 1997) composé de 12.069 occurrences.

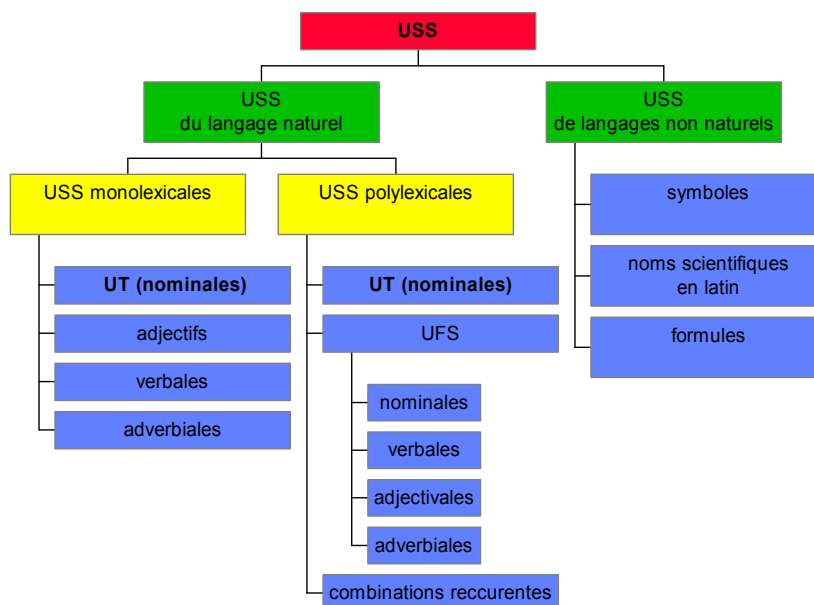
Les résultats des dépouillements manuels faits par les spécialistes sont les suivants :



1. Résultats des dépouillements manuels

L'analyse des résultats des dépouillements manuels renforce l'idée que l'UT n'est pas la seule unité de textes spécialisés qui transmet un signifié spécialisé pertinent. En accord avec cette nouvelle perspective, l'unité qui devrait être l'objet de dépouillement et, concrètement, l'objet d'un videur automatique, ne peut se réduire ni à l'UTP ni à l'UT, comprises comme des unités de catégorie grammaticale nominale. Au contraire, cet objet doit inclure toutes les unités de signification spécialisée (USS) des textes, aussi bien les USS de catégories grammaticales différentes qui font partie du langage naturel, que les unités qui font partie des langages non naturels; et, à l'intérieur des unités du langage naturel, cet objet devrait considérer les UT simples comme les complexes, les noms comme les verbes, les adjectifs et les adverbes, les unités lexicales comme les unités phraséologiques spécialisées (UFS); et, finalement, en ce qui concerne les unités de systèmes artificiels, il devrait comprendre aussi bien les symboles nominaux et les noms latins propres de nomenclatures consensées que les formules complexes.

Synthétisons ensuite le schéma des USS que nous proposons comme l'objet de travail d'une application terminologique, par exemple, d'un extracteur:



2. Typologie des USE

2.1 Les USS du langage naturel

La classification des unités des dépouillements de textes spécialisés présentée dans le graphique 1 permet de faire une première différenciation : les USS du langage naturel et celles de langages non naturels. Dans les premières, on peut distinguer les unités monolexicales (*mano, hipertensión, nervio, pleural, inmonológicamente, linfografía, vacunar*³) et les polylexicales (*enfermedad de Parkinson, afectación vascular, nervio renal posterior de Walter, nervio auditivo, aumento de la dosis de penicilina*⁴).

Du point de vue de la catégorie grammaticale, les USS monolexicales peuvent être des noms (*pie, vacunación, quimioterapia, laringotomía*⁵), des verbes (*inyectar, vacunar, desinfectar, infectar, hidratar*⁶), des adjectifs (*mononéfrico, cutáneo, nasal, alveolar*⁷) ou des adverbes (*clínicamente,*

³ En français: *main, hypertension, nerf, pleural, immunologiquement, lymphographie, vacciner.*

⁴ En français: *maladie de Parkinson, affection vasculaire, nerf rénal postérieur de Walter, nerf auditif, augmentation de la dose de pénicilline.*

⁵ En français: *pie, vaccination, chimiothérapie, laryngotomie.*

⁶ En français: *injecter, vacciner, désinfecter, infecter, hydrater.*

⁷ En français: *mononéphrique, cutané, nasal, alvéolaire.*

*immunológicamente, radiológicamente, biológicamente*⁸). Les USS polylexicales peuvent être des UTP, et dans ce cas, elles sont toujours des noms (*diagnóstico clínico, insuficiencia cardíaca grave*⁹), ou des unités phraséologiques (*tratar la hepatitis, aumento de la permeabilidad vascular*¹⁰). Il y a aussi des combinaisons non lexicalisées, mais discursivement récurrentes (*radiografía del pie izquierdo, diagnóstico de anemia*¹¹).

Morphologiquement, en médecine, les USS monolexicales nominales pertinentes peuvent être des unités simples (*fiebre, mano, hueso, pie*¹²), des unités dérivées (*nasal, enfebrarse, febril*¹³), des composés patrimoniaux (*cuentagotas, portaalgodón*¹⁴) ou des composés savants (*tifus, mieloma, pericardio, nefritis, cirrosis, anemia*¹⁵). Par opposition, les USS monolexicales verbales sont surtout des dérivées (*desinfectar, cicatrizar, vacunar*¹⁶) et les USS monolexicales adverbiales (*clínicamente, inmunológicamente*) ou adjectivales (*alérgico, clínico, medular, óseo*¹⁷), en biomédecine, ont toujours été construites par dérivation.

À côté des unités monolexicales et des polylexicales, et entre les unes et les autres, on peut placer un type d'unités qui sont aussi pertinentes: les sigles et les acronymes (*ADN, LTH, LTT, MAO, SIDA*). Du point de vue morphosyntaxique, ils sont des unités simples qui procèdent d'une concaténation d'unités, et qui, fonctionnellement, se comportent comme les substantifs car ils proviennent de syntagmes nominaux.

2.2 Les USS des langages non naturels

Au delà des USS du langage naturel, un extracteur devrait aussi pouvoir reconnaître les USS de langages non naturels des textes spécialisés, c'est à dire de langages construits par consensus. En médecine¹⁸, par exemple, les USS qui appartiennent à des nomenclatures artificielles se réduisent

⁸ En français: *cliniquement, immunologiquement, radiologiquement, biologiquement.*

⁹ En français: *diagnose clinique, insuffisance cardiaque grave.*

¹⁰ En français: *traiter l'hépatite, augmentation de la perméabilité vasculaire.*

¹¹ En français: *radiographie du pied gauche, diagnose d'anémie.*

¹² En français: *fièvre, main, os, pied.*

¹³ En français: *nasal, enfiévrer, fébrile.*

¹⁴ En français: *compte-gouttes, porte-cotton.*

¹⁵ En français: *typhus, myélome, péricarde, néphrite, cirrhoses, anémie.*

¹⁶ En français: *désinfecter, cicatriser, vacciner.*

¹⁷ En français: *allergique, clinique, médullaire, osseux.*

¹⁸ Chaque domaine de spécialité priorise un type d'unités non naturels différent.

fondamentalement aux noms scientifiques en latin, aux symboles et aux formules, et, plus spécifiquement, aux unités suivantes: les noms latins: des parties du corps qui constituent la Nomina anatomica (*arteria femoralis, vena femoralis, lobus anterior, lobus inferior*); des zoonims (*Rattus rattus, Diphylobothrium latum*); des plantes, aussi bien les comestibles que les vénéneuses et les médicinales (*Melissa officinalis, Artemisia herba-alba*); et des bactéries (*Mycobacterium tuberculosis, Rickettsia australis, Rickettsia conorii*); les symboles des éléments ou des composés chimiques, organiques et inorganiques (*Ra, Na, F, B₁, B₂, B₃*); du Système International des Unités (*L, J, s, A*) et les formules des composés chimiques (*CH₂O, H₂O, C₈H₁₀N₂S*).

2.3 En synthèse

Pour synthétiser, nous présentons une table avec les USS qu'un extracteur qui se propose d'obtenir un niveau d'exhaustivité élevée devrait récupérer à partir des textes du domaine des sciences de la santé. Nous avons encadré les unités que les extracteurs ont l'habitude de détecter, et nous avons souligné les USS référentielles nominales, c'est à dire les UT:

Unités de signification spécialisée (USS)

1. USS du langage naturel

1.1 USS linguistiques monolexicales

1.1.1 simples

1.1.1.1 nominales

1.1.1.2 verbales

1.1.2 dérivées

1.1.2.1 nominales

1.1.2.2 verbales

1.1.2.3 adjectivales

1.1.2.4 adverbiales

1.1.3 composées patrimoniales nominales

1.1.4 composées savantes nominales

1.1.5 sigles

1.2 USE linguistiques polylexicales

1.2.1 unités terminologiques polylexicales (UTP)

1.2.2 unités récurrentes nominales

1.2.3 unités phraséologiques spécialisées (UFS)

1.2.3.1 nominales

1.2.3.2 verbales

1.2.3.3 adjectivales

1.2.3.4 adverbiales

2. USE de langages non naturels

2.1 symboles

2.2 noms scientifiques en latin

En résumé, ce test indique que les USS de catégorie grammaticale nominale sont très nombreuses, mais qu'elles ne sont pas les seules. Cette constatation prouve que l'objet des applications terminologiques peut être plus vaste que celui des UT classiques. Malgré cela, ce premier test expérimental a été réalisé avec un groupe homogène d'utilisateurs de terminologie (des spécialistes) ; ainsi, en suivant le principe d'adéquation, si on diversifie les besoins professionnels, l'objet doit aussi varier. Ce point sera soulevé dans la dernière partie du travail.

3. Objet d'extraction et besoins professionnels

Dans la partie précédente nous avons remis en question le fait que la seule unité d'intérêt d'un texte spécialisé est l'UT, comprise comme une unité lexicale de caractère nominal, et, plus spécifiquement, l'UTP; pour cette raison et par opposition aux approches plus classiques et restrictives, nous avons substantiellement ouvert l'objet de travail des applications terminologiques.

Dans ce paragraphe, nous nous demandons si les USS sont les mêmes pour tous les usagers ou si, au contraire, selon les finalités professionnelles il y a un ensemble d'unités pertinentes et un ensemble non pertinentes. Ainsi, si dans le point antérieur nous nous sommes interrogés sur le quoi (l'objet d'étude), maintenant nous nous demandons le pour qui? (les usagers des produits terminologiques) et surtout le dans quel but? (les besoins terminologiques des activités professionnelles).

Nous voulons montrer comment les unités de signification spécialisée (USS) varient qualitativement et quantitativement selon les besoins professionnels dans lesquelles elles s'utilisent. Nous partons, donc, de l'hypothèse selon laquelle toutes les USS qu'il y a dans un texte ne sont pas pertinentes pour une activité professionnelle spécifique. Par conséquent, les USS d'un texte qui intéressent un spécialiste, un traducteur, un terminographe ou un documentaliste peuvent ne pas converger du tout.

Ainsi, en continuant avec l'exemple du dépouillement automatique, si on applique un extracteur qui ne se prend pas en compte le point de vue de l'utilisateur —comme le font la majorité—, il produit toujours la même sélection de termes. Mais, si on veut que les systèmes d'extraction puissent faire des sélections adéquates aux besoins professionnelles des différents groupes d'utilisateurs, on doit pouvoir profiler préalablement ces besoins.

Un test réalisé (Estopà, 1999) a démontré que chaque groupe professionnel choisit son propre ensemble d'unités pertinentes. Cette sélection inclut les unités sélectionnées, le type le plus représentatif et les paramètres qui limitent ou donnent la priorité à chaque différent choix.

Le test expérimental reposait sur l'hypothèse que des groupes professionnels différents, lorsqu'ils réalisent une activité professionnelle déterminée, s'approchent des textes spécialisés avec des intérêts différents. Ils ont donc des points de vue différents sur les USS pertinentes d'un texte spécialisé.

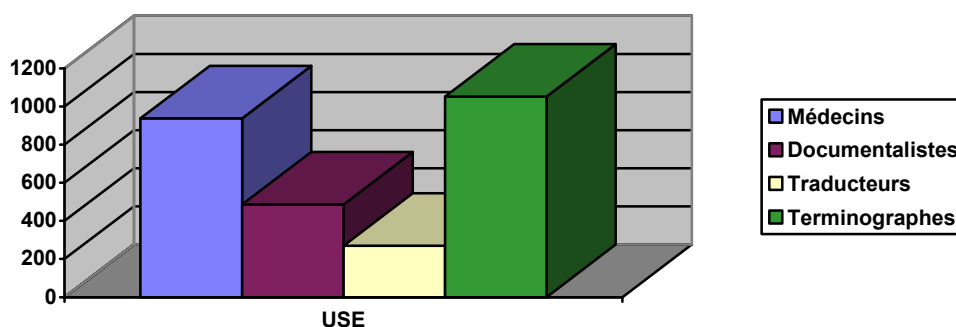
L'expérience consistait à réaliser le dépouillement des unités spécialisées pertinentes pour différentes activités professionnelles. Concrètement, nous avons sélectionné quatre groupes professionnels en rapport avec des textes spécialisés: spécialistes, documentalistes, traducteurs spécialisés et linguistes/terminographes. Le dépouillement du texte a été fait par trois spécialistes de chacun des groupes professionnels proposés.

En accord avec l'hypothèse selon laquelle c'est l'activité professionnelle, et non le collectif, l'élément pertinent qui conditionne la sélection des USS d'un texte, nous avons limité les activités

professionnelles à une seule finalité différente pour chaque groupe. Nous n'avons pas pris en compte qu'un même professionnel peut faire plusieurs activités à partir d'un texte de spécialité. En ce sens là, nous avons préféré nous fixer sur les activités professionnelles les plus prototypiques de chaque groupe: la transmission de la connaissance (des médecins), l'indexation de textes (des documentalistes), la traduction (des traducteurs spécialisés), et l'élaboration de terminologies (des terminographes).

Le texte que nous avons utilisé pour le dépouillement —Maladies produites par Rickettsia— est le même que celui de la première expérience.

Les résultats de ce test expérimental ont montré que toutes les USS d'un texte ne sont pas pertinentes pour toutes les activités professionnelles ; dans un texte spécialisé il y a des USS pertinentes pour tous les groupes, des USS pertinentes seulement pour quelques groupes, et des USS pertinentes uniquement pour un des groupes. Les données globales des quatre groupes professionnels renforcent l'idée que chaque collectif a son critère de sélection d'USS pertinentes pour réaliser son activité professionnelle. Cette diversification de critères implique, comme on peut le voir dans le graphique 3, une diversité au niveau du nombre d'USS sélectionnées:



3. Nombre total d'USS sélectionnées par chaque collectif professionnel

En effet, le graphique 3 démontre que le nombre d'USS sélectionnées ne converge pas selon les groupes. Les médecins ont signalé 938 unités différentes, les traducteurs 270, les documentalistes 486 et les terminographes 1.052. De plus, les quatre groupes montrent très peu de convergence quant aux unités

qu'ils ont marquées : seulement 9,3%. Ainsi, des médecins, documentalistes et terminographes seulement ont convergé dans 30,7% d'unités, et des médecins, traducteurs et des terminographes 11,3%. Les coïncidences augmentent considérablement si on compare les groupes deux par deux: des médecins avec des linguistes convergent dans 88%, des médecins avec des traducteurs dans 30%, des médecins avec des documentalistes dans 50%.

Cette constatation vérifie la supposition que, du point de vue fonctionnel, les USS (et également les termes) d'un domaine ou d'un objet thématique ne sont pas préfixées, sinon qu'elles changent selon les activités professionnelles. Ceci signifie que dans un même texte, il y a des mots qui sont uniquement pertinents pour un groupe en accord avec l'activité professionnel réalisée.

Autre caractéristique, que l'on peut observer dans le tableau 4 et qui découle de l'analyse globale des dépouillements: l'absence de convergence entre groupes en ce qui concerne les types d'USS soulignées:

	médecins		documentalistes		traducteurs		terminographes	
	Num.	%	Num.	%	Num.	%	Num.	%
Noms	824	87,84	426	87,65	211	78,14	900	85,56
Verbes	17	1,81	0	0	1	0,37	49	4,65
Adjectifs	28	2,98	5	1,02	27	10	35	3,32
Adverbes	5	0,53	0	0	2	0,74	4	0,37
Sigles	13	1,38	12	2,46	5	1,85	12	1,13
Symboles	6	0,63	3	0,61	0	0	6	0,56
Noms scientifiques	44	4,69	22	4,52	0	0	45	4,27
Noms propres	0	0	18	3,70	1	0,37	0	0
Phrases discursifs	1	0,10	0	0	23	8,51	1	0,09
Total	938	100	486	100	270	100	1052	100

4. Types d'unités sélectionnées par chaque collectif

En ce sens, nous observons dans le tableau 4 une gamme de possibilités qui va s'étendre de la considération des USS nominales (comme c'est le cas de quelques documentalistes) à la sélection de tous les types d'USS possibles (les choix des médecins). En général, les terminographes, suivis des spécialistes, sont ceux pour lesquelles, on a noté le plus d'unités et le plus de variétés en ce qui concerne la structure et la catégorie grammaticale. Les documentalistes et les traducteurs, par opposition, sont ceux qui ont souligné le moins d'unités.

En relation avec les aspects quantitatifs des dépouillements des docteurs, nous pouvons dire qu'ils ont souligné beaucoup d'unités, même si ce fait n'est pas toujours systématique. Dans l'application des critères de sélection, semble-t-il qu'ils tendent à être très systématiques dans la sélection des UT, des USS adverbiales, des sigles, des symboles, des noms scientifiques qui font partie de nomenclatures consensuées et des UT avec un noyau déverbal; et ils sont moins systématiques avec les UFS formées par des combinaisons très fréquentes de deux unités terminologiques comme traitement de + une maladie, radiographie de + art + partie anatomique; et ils sont très peu systématiques avec les USS adjectivales et verbales. Finalement, nous remarquerons qu'ils ne soulignent pas les UFS verbales ni les adverbiales, ni les noms propres isolés.

Mise à part l'hétérogénéité dans la sélection de quelques unités, les types d'USS que les spécialistes considèrent pertinentes, car elles sont des unités qui transmettent la connaissance spécialisée, sont les suivantes: des USS nominales, c'est à dire des termes, des USS verbales, des USS adjectivales, des USS adverbiales, des noms scientifiques, des symboles, des sigles, des UFS nominales et verbales, et des combinaisons récurrentes nominales.

En ce qui concerne les documentalistes, ils ont sélectionné très peu d'USS du texte par rapport aux autres informateurs et les types d'unités sélectionnées, comme nous le montre le tableau 3, se réduisent aux suivants: des unités monolexicales nominales (42,91%), des unités polylexicales nominales (55,85%), des adjectifs (0,41%), des sigles (2,46%), des symboles (0,61%), des noms scientifiques (mais seulement les nombres de micro-organismes) (4,51%) et des noms propres (3,69%). Ils considèrent non pertinentes ni les adverbes, ni les verbes et les UFS non plus. A la différence des informateurs restants, deux documentalistes ont souligné comme des unités valides pour indexer un texte des noms propres. Ces noms font référence à des pays dans lesquels on trouve les conditions qui provoquent une maladie

déterminée, à des écoles de médecine ou à des médecins célèbres, et ils facilitent la délimitation plus précise des recherches potentielles.

La fréquence d'utilisation des USS pertinentes et leur emplacement dans le texte sont deux données très significatives pour les documentalistes. En ce sens, si une USS figure dans un texte avec une fréquence élevée (c'est à dire qu'elle se répète dans chaque paragraphe ou dans chaque alinéa) et/ou elle fait partie du titre du document, de quelques sous-titres, des schèmes ou des sommaires, alors il existe une probabilité très élevée pour qu'elle soit une unité représentative du contenu du texte. En outre, les documentalistes tendent à sélectionner les unités les plus spécifiées (les UTP), parce qu'elles permettent d'avoir une plus grande précision. Le bruit occasionné par les unités très génériques ou polysémiques en est alors réduit.

Ainsi, nous pouvons affirmer que le dépouillement des USS pertinentes pour indexer des textes est différent des autres finalités professionnelles, en fonction des éléments suivants:

- la réduction des catégories grammaticales: seulement les nombres et quelques adjectifs
- le nombre peu important d'USS: seulement les unités représentatives du texte (en ce sens, la fréquence et la situation des unités dans le texte sont des données indispensables)
- l'importance que certains noms propres ont pour identifier un document
- la prédominance d'unités polylexicales.

Quant aux traducteurs, ils représentent le groupe qui a sélectionné le moins d'unités comme on peut le voir dans le tableau 3. Plus précisément, ils ont seulement sélectionné les unités qu'ils ne connaissent pas sémantiquement ou celles qui peuvent leur occasionner des problèmes de traduction. Pour cette raison, il y a une réduction considérable des types d'USS sélectionnées: ils ne sélectionnent pas de symboles, de noms scientifiques en latin, et ils sélectionnent très peu de sigles.

Cette restriction est assez logique, si on prend en compte le fait que les symboles et les noms en latin des nomenclatures scientifiques sont, en général, universels et qu'ils ne sont donc pas l'objet de traduction. Les sigles qui s'utilisent en médecine apparaissent normalement en anglais, car cette langue est aujourd'hui reconnue comme langue internationale et de ce fait élimine tous les problèmes de traduction. Cependant, les traducteurs ont sélectionné quelques sigles du texte les moins connus, même s'ils ne se traduisent pas. De plus, les sigles ont toujours été soulignés avec leur contexte référentiel.

Autre particularité du dépouillement des traducteurs est le fait qu'ils ont souligné les USS dans le contexte syntaxique car, à plusieurs reprises, c'est le contexte qui fournit des éléments linguistiques et pragmatiques pour proposer les équivalents le plus adéquats. Ils ont aussi sélectionné les référents socioculturels du texte qu'ils devront adapter ou expliquer lors de la traduction.

Finalement, les terminographes sont les plus exhaustifs dans la sélection d'unités, même s'ils ne profitent pas de toutes les USS sélectionnées pour élaborer un dictionnaire concret.

Il est intéressant d'observer que, quand les terminographes soulignent des verbes, ils se rapportent aussi à leur contexte d'utilisation. Le contexte peut faciliter des exemples, mais aussi il peut indiquer la présence de phraséologie verbale. Cette phraséologie a été traditionnellement exclue des dictionnaires spécialisés, mais, lors des dernières années, on a commencé à considérer la pertinence de l'introduire dans ce type d'ouvrage.

3.1 En synthèse

L'analyse des résultats des dépouillements de plusieurs groupes permet de valider l'hypothèse formulée au début et qui défend la non correspondance dans la sélection des USS d'un texte entre les groupes professionnels par rapport à deux paramètres: le nombre d'unités et le type d'unités.

En effet, les données démontrent que les différences de dépouillements entre les professionnels sont très significatives aussi bien au niveau du nombre d'unités qu'ils sélectionnent que des types d'USS sélectionnées. Cette divergence s'explique par le fait que chaque activité professionnelle requiert des USS précises pour réaliser leurs propres fonctions et, à la fois, elles ne tiennent pas en compte les autres qui auraient pu être pertinentes pour une autre activité.

Mais, que la sélection d'USS soit spécifique pour chaque activité ne présuppose pas que le concept d'USS soit pluriel. Depuis notre point de vue, la notion d'USS est unique et ce qui change c'est le concept d'USS pertinente, de manière qu'une USS peut être non pertinente pour une activité. Les analyses des résultats obtenus confirment et valident l'hypothèse que la pertinence d'une USS dépend de la finalité professionnelle.

Ainsi, pour les spécialistes en médecine, les USS pertinentes sont toutes les unités qui transmettent la connaissance spécialisée, et cette condition la vérifie, par définition, toutes les USS d'un

texte. Pour les médecins, les USS sont un sous-ensemble des unités de cognition spécialisé. Pour cette raison, le dépouillement des médecins est plus exhaustif autant en ce qui concerne la quantité que la diversité des types d'unités.

Pour les terminographes, dans une première étape toutes les USS du texte sont aussi pertinentes, ce qui explique la coïncidence de sélections entre les médecins et les terminographes. Mais, lors d'une deuxième étape, le terminographe sélectionne, entre toutes les USS sélectionnées, seulement celles qui sont adéquates aux objectifs, aux destinataires et aux fonctions de l'ouvrage terminographique.

Le schéma cognitif de pertinence que les documentalistes ont des USS d'un texte est plus restrictif que celui des spécialistes et des terminographes; pour cette raison, leur sélection d'USS est aussi plus réduite quantitativement et typologiquement.

Un documentaliste est seulement intéressé par les USS qui représentent plus précisément le contenu général du texte dans lequel elles apparaissent. Ainsi, après une analyse conceptuelle du document, il sélectionne les USS qui fonctionnent comme des étiquettes de son contenu informatif. Les USS pertinentes sont, donc, des mots d'identification qui permettent de décrire, d'indexer, d'ordonner et de récupérer un document. Cette affirmation explique qu'il souligne presque uniquement des noms, car ces derniers représentent les USS qui décrivent le plus synthétiquement le contenu d'un texte, et dans les noms, les unités polylexicales, parce qu'elles sont les unités qui permettent de limiter le plus de recherches futures.

Finalement, dans un niveau de restriction plus élevé du point de vue quantitatif, mais plus faible du point de vue typologique, on trouve les traducteurs. Les USS qu'ils soulignent, avant d'aborder une traduction, sont seulement celles qui peuvent occasionner des problèmes de traduction. Les difficultés que les USS posent peuvent être cognitives, linguistiques ou socio-fonctionnelles, et l'analyse de leurs dépouillements confirme qu'ils sont seulement intéressés par les unités dont ils ne connaissent pas le signifié ou bien par les unités supposées problématiques. Pour cette raison, parmi les résultats de leurs dépouillements nous trouvons aussi des segments d'USS (et pas des unités entières), surtout des adjectifs et des noms non spécialisés qui ont l'habitude d'intégrer des unités polylexicales. D'un autre côté et en accord avec cette logique, toutes les USS qui ne se traduisent pas ne sont pas considérées (comme les symboles, les noms qui font partie de nomenclatures, quelques sigles).

4. Profils de besoins différents

Les analyses quantitatives et qualitatives des dépouillements manuels de différents groupes avec des besoins professionnels concrets nous permettent de constituer des profils qui annoncent, d'un côté, les types d'unités à prendre en compte et, de l'autre, des informations sur ces unités comme la fréquence d'usage, le contexte d'utilisation ou la situation dans le texte.

En accord avec ces deux variables, nous avons établi les quatre profils par rapport aux activités professionnelles suivantes: la transmission de la connaissance spécialisée (spécialistes), l'indexation d'un texte (documentalistes), la traduction spécialisée (traducteurs spécialisés) et la pratique terminographique (terminographes).

Pour les spécialistes, les USS pertinentes d'un texte de leur spécialité sont toutes les USS du texte, mais spécialement les noms car ils concentrent et représentent mieux la connaissance.

Pour les spécialistes en documentation, les USS pertinentes d'un texte de spécialité, comme nous l'avons déjà vu, sont celles qui fonctionnent comme des étiquettes du contenu informatif du texte et qui permettent de décrire, d'indexer, d'ordonner et de récupérer un texte spécialisé déterminé. En conséquence, il serait très utile, lorsque l'on utilise un extracteur pour indexer un texte, qu'il fournisse l'information sur la fréquence et sur la disposition discursive des USS dans le corpus textuel, de manière à montrer seulement les USS (préférentiellement des unités nominales polylexicales) qui sont supérieures à une fréquence déterminée, accompagnées avec l'information de leur situation dans le texte.

Les USS qui intéressent les traducteurs sont seulement celles qui peuvent poser certaines difficultés au moment de les traduire. Pour cette raison, ils sélectionnent parfois seulement des segments d'UTP (et pas l'unité entière), surtout les unités (nominales ou adjectivales) non spécialisées qui intègrent quelques unités polylexicales. Le fait que chaque traducteur ait des besoins cognitifs, linguistiques et socio-fonctionnels différents, qui dépendent de son niveau de connaissance du thème dans les deux langues, implique qu'il n'y ait pas de types d'unités prototypiques qui les intéressent plus que d'autres. Tout dépend de leur expérience professionnelle.

Malgré cela, nous sommes aussi arrivés à la conclusion que les USS qui occasionnent le plus de problèmes de traduction sont: les UFS, les sigles non internationalisés, les éponymes, les USS ou les segments d'USS non spécialisés et les néologismes. Par conséquent, si un extracteur doit être utilisé pour

les besoins terminologiques de la traduction, il est intéressant qu'il puisse récupérer toutes les unités linguistiques avec leur contexte d'utilisation, car plusieurs fois le contexte offre des données qui facilitent la compréhension de l'unité ou la recherche de son équivalent.

Et finalement, pour les terminographes, il serait important qu'un extracteur récupère toutes les USS du texte avec leur contexte et leur fréquence d'utilisation, et avec une relation des USS linguistiques et non linguistiques.

5. Conclusions

En premier lieu, nous avons mis en relief la nécessité d'ouvrir l'objet de travail en terminologie et plus concrètement celui des extracteurs qui sont des outils qui aident les usagers dans leur travail terminologique. Nous avons également postulé le concept d'unité de signification spécialisée qui permet de considérer comme objet d'étude d'autres unités utilisées avec une valeur spécialisée, catégoriellement, syntaxiquement et sémantiquement différentes des termes.

Dans un deuxième temps, nous avons confirmé l'idée que la pertinence d'une USS dépend des besoins professionnels qu'une activité génère, en établissant les besoins terminologiques de quatre activités qui impliquent l'usage de textes spécialisés: la transmission de la connaissance, l'indexation, la traduction et l'élaboration de dictionnaires.

Ainsi, nous sommes arrivés à la conclusion que la notion d'USS est la même pour tous les groupes professionnels, tandis que la notion d'USS pertinente peut changer. Entre ces deux concepts, il y a une différence de restriction conditionnée par la fonctionnalité. Par conséquent, nous avons établi les besoins terminologiques des quatre activités professionnelles analysées et nous avons observé que certains d'entre elles, avec la finalité de faciliter la sélection définitive, requièrent que les USS soient accompagnées d'information complémentaire relative à leur contexte d'usage immédiat, à leur fréquence et/ou à leur situation dans le texte.

Avec cette deuxième expérience, nous avons voulu démontrer ceci: si on veut que les applications terminologiques soient efficaces et en adéquation avec leur usage, on doit prendre en compte les besoins professionnels que génère une activité déterminée. En ce sens, il nous semble utile de distinguer l'unité de base de la terminologie, de l'unité pertinente pour un travail terminologique concret.

Finalement, nous avons proposé quatre profils de besoins terminologiques qui génèrent les textes de spécialité. Des profils qui répondent aux nécessités les plus généralisables de chaque activité étudiée. À mesure que l'on avancera dans l'analyse de chaque activité, on pourra, d'une part, construire des profils nouveaux, et, d'autre part, proposer des éléments qui permettent d'affiner ces profils.

6. Bibliographie

AMAR, M. et DAVID, S. 2001. « Évaluation de logiciels d'extraction dans les champs de l'indexation, la traduction et la terminologie ». Corpus INRA. Rapport établi dans le cadre de l'ARC A2 (AUF).

AUGER, P. et ROUSSEAU, L. 1978 (1984). *Metodologia de la recerca terminològica*. Barcelona : Generalitat de Catalunya.

BOURIGAULT, D. 1993. "Analyse syntaxique locale pour le repérage de termes complexes dans un texte". *TAL*, 2, 105-117.

BOURIGAULT, D. 1994. *LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquisition des connaissances à partir de textes*. Paris : École des Hautes Études en Sciences Sociales. [Thèse doctorale].

CABRÉ, M. T. (dir.). 1996. *Terminologia. Selecció de textos d'E. Wüster*. Barcelona : Servei de Llengua Catalana-Universitat de Barcelona.

CABRÉ, M.T. 1992 *La terminologia. La teoria, els mètodes, les aplicacions*. Barcelona : Empúries.

CABRÉ, M.T. 1999 *La terminologia. Representación y comunicación. Una teoría de base comunicativa*. Barcelona: IULA, Universitat Pompeu Fabra. (Sèrie Monografies, 3).

CABRÉ, M. T. 2000 "Do we need an autonomous theory of terminology?". *Terminology*, 5, 1, 1998/1999, 2-20.

CABRÉ, M.T. (ed.). 2001. *La terminologia científico-técnica*. Barcelona: IULATERM, Universitat Pompeu Fabra.

CABRÉ, M. T. et ESTOPÀ, R. "On the units of specialised meaning uses in professional communication". *ITTF*, 2001, 1. (sur presse)

CABRÉ, M. T. ; ESTOPÀ, R. et VIVALDI, J. 2001 "Automatic term detection: a review of current systems". Dans: D. Bourigault, C. Jacquemin, M.-C. L'Homme (eds.) (2001) *Recent Advances in Computational Terminology*. Amsterdam: John Benjamins.

- CABRÉ, M. T. et FELIU, J. (eds.). 2001 *La terminología científico-técnica*. Barcelona : IULATERM, Universitat Pompeu Fabra.
- CONDAMINES, A. 1995. "Terminology: new needs, new perspectives". *Terminology*, 2, 2, 219-238.
- DAGAN, I. et CHURCH K. 1994. "Termight: Identifying and translating technical terminology". Actes de la *Fourth Conference on Applied Natural Language Processing*. Stuttgart.
- DAILLE, B. 1994. *Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques*. Paris : Université Paris VII. [Thèse doctorale].
- DAILLE, B. et al. 1996. "Empirical observation of term variations and principles of their description". *Terminology*, 3, 2, 197-257.
- DAVID, S. 1995. *Les Unités nominales polylexicales. Éléments de description et reconnaissance automatique*. Paris : Université Denis Diderot, Paris VII. [Thèse doctorale].
- DAVID, S. et PLANTE, P. 1991. "Le progiciel TERMINO: de la nécessité d'une analyse morphosyntaxique pour le dépouillement terminologique des textes". Actes du Colloque de Montréal *Les industries de la langue: perspectives des années 1990*, 1991, 1, 71-88.
- DROUIN, P. 1997. "Une méthodologie d'identification automatique des syntagmes terminologiques: l'apport de la description du non-terme". *Meta*, XLII, 1, 45-54.
- DUBUC, R. 1985. Manuel pratique de terminologie. Montreal : Linguattech.
- ENGUEHARD, C. et PANTERA, L. 1994. "Automatic Natural Acquisition of a Terminology". *Journal of Quantitative Linguistics*, 2, 1, 27-32.
- ESTOPÀ, R. 1999. Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'extracció automàtica de candidats a unitats de significació especialitzada). Barcelona: Universitat Pompeu Fabra. [Thèse doctorale].
- ESTOPÀ, R. 1999 "Eficiencia en la extracción automática de terminología". *Perspectives: Studies in Traductology*, 7, 2, 277-286
- ESTOPÀ, R.; VIVALDI, J et CABRÉ, M. T. (2000) "Extraction of monolexical terminological units: requirement analysis". *Second International Conference on Language Resources and Evaluation: Terminology Resources and Computation Proceedings*. Atenes: National Technical University of Athens, II, 51-56.

- EVANS, D. A. et ZHAI, C. 1996. "Noun-phrase Analysis in Unrestricted Text for information retrieval". Actes del 34th Annual Meeting of ACL. Santa Cruz: University of California, 1996, 17-24.
- FRANTZI, K. et ANANIADOU, S. (1995). "Statistical measures for terminological extraction". Manchester : *Working Papers du Department of Computing of Manchester Metropolitan University*.
- GUILLET, A. 1990. "Reconnaissance des formes verbales avec un dictionnaire minimal". *Langue française*, 87, 52-58
- HEID, U. et al. 1996. "Term extraction with standard tools for corpus exploration. Experience from German". *TKE '96: Terminology and Knowledge Engineering*. Berlin: Indeks Verlag, 139-150.
- JACQUEMIN, C. 1994. "Recycling Terms into a Partial Parser". Actes de la 4th Conference on Applied Natural Language (ANLP'94). Stuttgart, 113-118.
- JACQUEMIN, C. 1996. "What is the tree we see through the window: A linguistic approach to windowing and term variation". *Information Processing & Management*, 32, 4, 445-458
- KAGEURA, K. et UMINO, B. 1996. "Methods of Automatic Term Recognition". Working Papers: *National Center for Science Information Systems*, 1-22.
- LAURISTON, A. 1994. "Automatic recognition of complex terms: Problems and the TERMINO solution". *Terminology*, 1, 1, 147-170.
- LERAT, P. 1995. *Les langues spécialisés*. Paris : PUF.
- L'HOMME, M. 1996. "Sélection des prépositions dans les termes complexes Nom (Prép.) Nom à partir de leur structure conceptuelle". *Cahiers de Lexicologie*, 68, 1, 25-43.
- NAULLEAU, E. 1998. Apprentissage et filtrage syntatico-sémantique des syntagmes nominaux pour la recherche documentaire. Paris : Université Paris VIII. [Thèse doctorale].
- OTMAN, G. 1991. "Des ambitions et des performances d'un système de dépouillement terminologique assisté par ordinateur". *La banque des mots*, 4, 59-96.
- PERRON, J. 1989. "Termino: un système de dépouillement terminologique". *Terminogramme*, 54, 3-9.
- REY, A. 1979. *La terminologie: noms et notions*. (Que sais-je? 2d.) Paris: Presses Universitaires de France.
- SAGER, J-C. 1990. *A Practical Course in Terminology Processing*. Amsterdam and Philadelphia: John Benjamins.

SLODZIAN, M. 1995. "Comment revisiter la doctrine terminologique aujourd'hui?". *Le banque des mots*, 7, 11-18.

VOUTILAINEN, A. 1993. "NPtool, a detector of english noun phrases". Actes del *Workshop on Very Large Corpora*. Columbus: Ohio State University.